Data and Code for: "Monthly food prices from artificial intelligence (AI) strongly correlates with high-frequency crowdsourced prices within a fragile context"

Julius Adewopo, Bo Pieter Johannes Andree, Helen Peter, Gloria Solano-Hermosilla, Fabio Micale

Data and Code Availability Statement

The paper uses public and non-confidential data on commodity food prices that was published by WordBank Development Data Group . The archive contains the data in the folder "data/". The file is called WB_monthly_data.csv. We downloaded the "Microdata" on October, 3, 2023 from https://microdata.worldbank.org/index.php/catalog/study/NGA_2021_RTFP_v02_M

The paper uses public, non-confidential data from the European Commission Joint Research Center (EC-JRC) on crowdsourced food prices in the northern region of Nigeria. The archive contains the data in the folder "data/". The file is named "fpca_all.csv".xlsx. We downloaded the "Post -sampled Weekly Price" on October, 3, 2023 from https://datam.jrc.ec.europa.eu/datam/mashup/FP_NGA/index.html?_r=1

The paper uses public, non-confidential data from the European Commission Joint Research Center (EC-JRC) on crowdsourced food prices in the northern region of Nigeria. The archive contains the data in the folder "data/". The file is named "Raw_groundref_FPCA_0km_fnl.csv". We downloaded the "Step 2. FPCA" data on October, 3, 2023 from https://data.jrc.ec.europa.eu/dataset/f3bc86b0-be5f-4441-8370-c2ccb739029e.

The IDs for the Enumerators are 857846, 999678, 715942, 1000142, 919020, 417510, 759919, 1000800, 382013, 1000136, 972481, 676454, 456. These IDs were assigned during data collection to differentiate volunteer citizens from trained enumerators for the purpose of data validation. All other IDs should be assigned to the Crowd.

Computational Requirements

Software and Hardware Requirements

-Software: R. We used Version 3.6.1, but other versions should work too, especially those >= 4.1.0.

You may also need to install Rtools 4.0: https://cran.r-project.org/bin/windows/Rtools/rtools40.html

-Packages: There are many of them. They are all embedded at the beginning of the code line in each script

-OS: We used Windows 10 Pro. Other versions of Windows, as well as Mac and Linux, should work too.

-CPU: We have Intel(R) Xeon(R) CPU E3-1545M v5@ 2.9GHz 4 Cores 8 Logical Processors

-Installed RAM: 64 GB

Downloading and opening the replication files

If you are cloning the repository from Github (https://github.com/englander/access), open RStudio, click File -> New Project -> Version Control -> Git, paste "https://github.com/englander/access.git", and click Create Project.

Data preparation and reformatting

To generate the final dataset used in the analysis from the raw data sources mentioned above, initial data preparation is required: The following steps should be taken for each data source –

Monthly AI Price Data
Start by querying the data columns "Adm1_name" (select: Kano, Kaduna, and Katsina), "Date" (select: 01/01/2019 to 12/01/2021), Create a new column to recalculate prices per kilogram for variables "o_maize", "h_maize", "l_maize", "c_maize", "o_rice", "h_rice", "l_rice", "c_rice". Create a new column for Commodity label and rearrange the headers to show all price at open (Price_O), high (Price_H), low (Price_L), and close (Price_C) in separate columns. This is the "WB_monthly_data.csv".

Intraday Groundtruth validation data for 8 months
Start by querying the "Dataset_Step_2" file with the IDs of trained Enumerators in the as stated above. Create a new column and assign "Enumerator" label to all data records that accrue to the IDs, other IDs should be assigned as "Crowd". Query the "level2code" column for focal states (i.e. "Kaduna", "Kano", and "Katsina"), and "submission_date" for 01-Mar-2021 to 30-Oct-2021. Query the "Product" column to select "indian_rice", "thailand_rice", "white_maize", and "yellow_maize" and transpose the columns so that the corresponding values in the column "price_kg" is assigned to each commodity. Create a new column named "month" and assign the month name that corresponds to the date. Retain relevant columns, including "market_type_cat", "lat", "lon" etc and rename them accordingly, as in the dataset.

Daily crowdsourced groundtruth data for 3 years
Start by filtering the "Post-sampled weekly Prices" file by selecting the inclusive years (2019-2021), create a new column to assign the beginning date of the week for each corresponding record associated with the "Week" column. Filter for relevant commodities by selecting within the column "Commodity" for "Maize (white)", Maize (yellow), Rice (Thailand), and Rice (Indian). Under the column "Crowdsourcing Stability Indicator", filter for records that have values ≥ 0.8. For the "Commodity" column, create a new column and label the header as "Commodity_type", then assign underscore (_) to the commodity and subtype e.g. Maize (white) should be assigned as "maize_white. Revise the records in "Commodity" column to show "maize" for all records that pertain to maize, and "rice" to records that pertain to Rice. Change the column named "Post_sampled price" to "Price" and column named "PRICE TYPE" to "Market". Remove non-relevant columns to mirror the "fpca_all.csv" dataset.

For the full details regarding the construction of all the specific variables used in the analysis, please refer to the paper.

For easier stepwise replication of analysis and outputs, 2 major scripts are presented and the code should be run in stepwise chunks, following the notes in the script

Analysis

Step 1:

Run the "Raw_crowd_ref_analysis_v1.R* scripts. The first code lines loads the 8-months validation dataset from your designated working directory. Modify the directory in the script to point to the right folder where the input data is saved.

Run Code lines 31 -288 for initial insight on overall trend of the data for each commodity daily, weekly, and monthly

Run Code lines 290 – 642 for assessment of relationship between crowd-submitted and enumerator-submitted groundtruth

Run Code lines 645 – 653 for Count of datapoints per Commodity;

Step 2: Run the "Final_WB_FPCA_Code.R"* script. The first code chunk loads the 3 year datasets from groundtruth and AI-estimates. Modify the directory in the script to point to the right working folder/directory where the input data is saved.

Run Code lines 1- 123 to calculate averages of groudtruth and AI-estimated prices on monthly basis, for Maize Commodity

Run code lines 130 -149 to implement region-wide analysis of the relationship and to assess coefficient of determination, for Maize Commodity

Run code lines 173 -208 to generate a table of relationship between prices from both data sources, disaggregated by State (Admin level2), for Maize commodity

Run code line 212 - 323 to analyze relationship by each market segment, for Maize Commodity

Run Code lines 348 - 412 to calculate averages of groudtruth and AI-estimated prices on monthly basis, for Rice Commodity

Run code lines 418 - 443 to implement region-wide analysis of the relationship and to assess coefficient of determination, for Rice Commodity

Run code lines 445 -500 to generate a table of relationship between prices from both data sources, disaggregated by State (Admin level2), for Rice commodity

Run code line 504 - 530 to analyze relationship by each market segment, for Rice Commodity

Figures

To reproduce the figures, execute the scripts in the following order and all figures should be generated in the specified working directory

1. Run the code line in "Raw_crowd_ref_analysis_v1.R" first;
2. Run the code in "Paper_charts.R" script,
3. Run the code lines in "validation_code_fnl_v1.R"
4. Run the code in "Final_WB_FPCA.R"

Note that Figure 1 was created in ArcMap so this cannot be reproduced with R script, while figure 2a was created in Excel.

The overall runtime is approximately 20 minutes;

The table in the paper was manually prepared by extracting the corresponding numbers for the metrics from analysis.